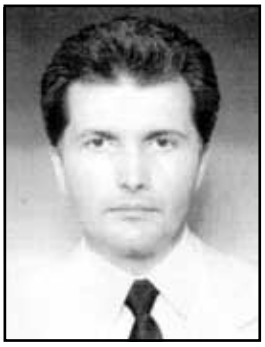
Using Property Based Sequence Motifs and 3D Modeling to Determine Structure and Functional Regions of Proteins

Ovidiu Ivanciuc†, Numan Oezguen†, Venkatarajan S. Mathura†, Catherine H. Schein, Yuan Xu and Werner Braun*



Sealy Center for Structural Biology, Department of Human Biological Chemistry and Genetics, University of Texas Medical Branch, Galveston, TX 77555-0857, USA

Abstract: Homology modeling has become an essential tool for studying proteins that are targets for medical drug design. This paper describes the approach we developed that combines sequence decomposition techniques with distance geometry algorithms for homology modeling to determine functionally important regions of proteins. We show here the application of these techniques to targets of medical interest chosen from those included in the CASP5 (Critical Assessment of Techniques for Protein Structure Prediction) competition, including the dihydroneopterin aldolase from *Mycobacterium tuberculosis*, RNase III of *Thermobacteria maritima*, and the NO-transporter nitrophorin from saliva of the bedbug *Cimex lectularius*. Physical chemical property (PCP) motifs, identified in aligned sequences with our MASIA program, can be used to select among different alignments returned by fold recognition servers. They can also be used to suggest functions for hypothetical proteins, as we illustrate for target T188. Once a suitable alignment has been made with the template, our modeling suite MPACK generates a series of possible models. The models can then be selected according to their match in areas known to be conserved in protein families. Alignments based on motifs can improve the structural matching of residues in the active site. The quality of the local structure of our 3D models near active sites or epitopes makes them useful aids for drug and vaccine design. Further, the PCP motif approach, when combined with a structural filter, can be a potent way to detect areas involved in activity and to suggest function for novel genome sequences.

Keywords: CASP5, MASIA, PCPMer, sequence motifs, physical-chemical properties, Bayesian statistics, functional annotation, drug and vaccine design.

INTRODUCTION

Rational design of drugs and vaccines is dependent on accurate 3D structural information for proteins. Protein modeling is particularly useful for predicting the effects of alterations in a protein of known sequence, and for extracting additional information from existing structures. For example, the models we prepared for human decay accelerating factor [1], measles virus receptor CD46 [2], pollen allergen Jun a 3 [3] and the mitochondrial cytochrome P450-27a1 [4] could be used to derive testable hypotheses about the function or epitope structure of these proteins. The best way to validate new structure prediction techniques is to compare models to subsequently determined experimental 3D structures. For example, we prepared a model of the CP1 and 2 domains of the CD46 protein, a receptor of the vaccine strain of measles virus, based on a template with only 20% identity [2]. Two years later, a crystal structure for CD46 was released, which differed from our model structure by only 1.6 Å (backbone root mean square deviation, bb-RMSD) [5]. Even more remarkable, the orientation of the two domains in the model was correct.

The CASP (Critical Assessment of Techniques for Protein Structure Prediction) competitions [6] were established as a rational way to compare the accuracy of methods now in use for modeling protein structures. The sequences of 63 proteins in CASP5, occasionally with some additional information, were provided to the ~200 participating groups, who were given 1-2 months to provide models of the protein structures. The models were then compared to the experimental 3D structures, released after the competition closed. Parts of the methods described here were developed during our participation in CASP4 [7]. We used our participation in CASP5 to test whether our motif recognition methods, described below, would aid in selecting among protein alignments generated by fold recognition servers.

Homology, or template based, modeling means that the structure of a novel protein is predicted based on the structure of a similar protein. As illustrated by the examples in this article and by other models submitted in the competition, accurate models can now be made even when the template and target protein are only distantly related, when the overall sequence identity is for example even <20%. Part of this success is due to the excellent fold recognition servers now available. Among the more powerful ways to enhance the quality of models at low levels of identity is to match common motifs that can be related to 3D structure [8, 9]. Sequence profile searches [10-15] and hidden Markov Models [16-18] can find distantly related proteins, but do not generally specify the critical locations in

*Address correspondence to this author at the Sealy Center for Structural Biology, Department of Human Biological Chemistry and Genetics, 301 University Boulevard, University of Texas Medical Branch, Galveston, TX 77555-0857, USA; Tel: (409) 747-6810; Fax: (409) 747-6000; Email: werner@newton.utmb.edu

†these authors contributed equivalently to the work detailed in this review.

0929-8673/04 $45.00+.00

the protein sequence that are relevant for structural or functional similarity. Conventional methods to detect common patterns of conserved residues, such as PROSITE [19] or BLOCKS [20], rely to a large extent on strictly conserved residues, and can miss subtle sequence motifs.

We have developed a sequence decomposition method to identify areas of conserved physical chemical properties (PCPs) in proteins that have low overall sequence identity, which is implemented on our MASIA website, [21], (http://www.scsb.utmb.edu/masia/masia.html), and in the stand alone program PCPMer [9]. We are now testing the use of these techniques, in combination with our previously developed modeling methods, to improve the quality of alignments for homology modeling. In this review we illustrate the application of these methods to selected targets from the CASP5 competition that are of interest to medical research groups. These comprise drug targets, including the bacterial enzymes RNase III, DHNA, and methionine aminopeptidase. We also illustrate how matching PCP motifs can be used to suggest the function of novel proteins revealed by the genome initiative.

GENERAL MODELING STRATEGY FOR CASP5

Collecting Diverse Sequences for Target Family Members Using BLAST

An overview of our general modeling strategy is illustrated in Fig. (**1**). Protein sequences related to the target sequence were identified with a BLASTP/PSIBLAST [22, 23] search in the non-redundant sequences (nr) available at NCBI. For PSIBLAST we set a maximum of five iterations with E-cutoff of 0.005 (or 0.001 for BLASTP). The taxonomy classification was used to select sequences at the genus level to insure that the sequences of family members were sufficiently diverse (the ideal alignment contains sequences that are between 25 and <80% identical). The sequence annotation is used to discard hypothetical or putative members, except when the target itself (e.g. T188) is hypothetical. A multiple alignment with the target sequence on top is then generated with CLUSTALW [24].

Physical-Chemical Property (PCP) Based Motif Detection for a Protein Family Using MASIA

We previously demonstrated that conservation of PCPs among sequences of protein families can be conveniently defined in terms of five physical-chemical vector components. These quantitative descriptors were derived from a large number (237) of PCPs, and decomposed into five principal components E_1-E_5 [25]. The five vectors were shown to sufficiently capture the distribution of amino acids in the original property space with an accuracy of 99%. Each of the five vectors is a linear combination of several properties. The first component E_1 correlates best with hydrophobicity. Conservation of these descriptors at a residue position in a sequence family indicates physical-chemical property conservation due to evolutionary constraints. Such conserved positions may not be detected by simply comparing the amino acid alphabet conservation. The reader is referred to the original paper for more details [25]. We demonstrated that our method can find distantly

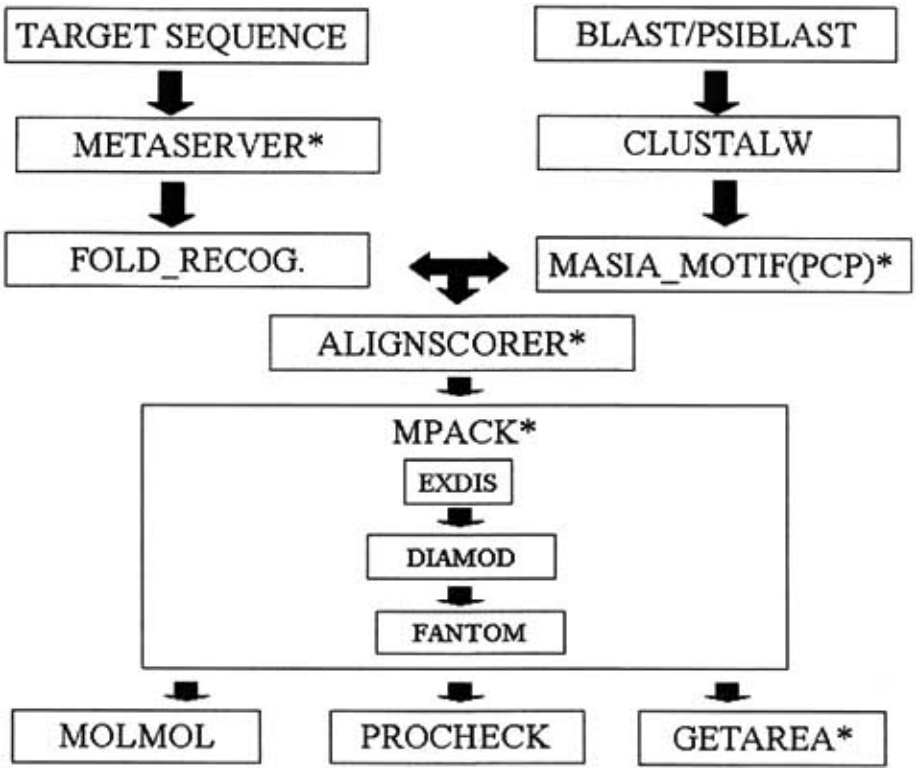


Fig. (1). Overview of our general modeling strategy, which incorporates physical chemical property (PCP) motif recognition to score alignments from fold recognition scorers. *indicates in-house software

related proteins to the DNA repair protein family APE [9] using property based sequence motifs. A Bayesian scoring system of the PCP motifs ranked all members of the DNase-I like SCOP-superfamiliy of APE as top scoring protein. Other high scoring proteins were from different SCOP classifications [26], but shared functions with the APE/DNase-I/IPP superfamily, including phosphatase activity and/or metal ion binding.

The ClustalW multiple alignment of the sequences from target family is used as input to the MASIA [21] (http://www.scsb.utmb.edu/masia/) program. The MASIA program is a tool to detect motifs in a sequence alignment either based on the identity of the amino acids, or conservation of the property vectors (E_1-E_5) described above [9]. At each position in the multiple alignment, MASIA calculates the average magnitude of the five vectors, standard deviation and relative entropy that constitutes a quantitative enumeration of motifs (called profiles). The relative entropy is calculated using the natural frequency of amino acid occurrence and the observed frequency distribution in the multiple alignment at a particular position. If the observed distribution is significantly different compared to the natural frequency, then the relative entropy will be high. Positions that have entropy value greater than a cut-off of 1.25 (default value in the program) are considered to be significant. A list of motifs are defined for a protein family using a G-cutoff, a parameter to filter out insignificant residue positions allowing gaps, and L-cutoff that limits the length of minimum size of a motif [9]. Each motif is quantitatively expressed as a profile and is subsequently used in alignment improvement and selection of template sequences.

TEMPLATE SELECTION AND ALIGNMENT IMPROVEMENT USING ALIGNSCORER

To determine suitable templates, the target sequence is submitted to several fold recognition servers (see Table 1), which return the alignments via email. For CASP5, most server results were available via the CAFASP website. Alignments of templates with targets above a certain cutoff score are collected and biological details, such as function and location of important residues are obtained from the literature. As there are often several PDB file names for the same protein, the list is sorted according to the SCOP classification of the template. In the easiest case, for targets that have a clear structural homolog in the PDB, all servers will list the same template as the highest scoring with identical alignment. Our program ALIGNSCORER (or CAFASPSCORER) is used to discriminate between alignments and templates. As Table 2, containing excerpts from the CAFASPSCORER results for target T184, demonstrates, the programs rank alignments from the servers according to the quality of fit between the template sequence and the target motif regions detected by MASIA. A Lorentzian-based additive scoring scheme that uses the motif profile and the template sequence measures the goodness of fit in the motif region [9]. We assume that the correct alignment with a truly homologous template will match all or most of the highly conserved motifs. The 3D-PSSM alignments (highlighted in Table 2) were selected for modeling the two domains of T184 (see modeling details below). Output for these alignments from ALIGNSCORER, a tool to determine motif scores for individual alignments, is illustrated in Table 3.

Three-Dimensional Model Generation for Proteins Using the Modeling Package (MPACK)

Once a suitable template and alignment is identified, we used our in house modeling suite MPACK to generate a model. MPACK combines EXDIS and DIAMOD to:

1. extract distances between the i^{th} and j^{th} atoms (d_{ij}), dihedral angles for the backbone (ψ, ϕ, Ω) and side chains (Chi) of matched residues from the template structure with the program EXDIS [27]. To allow flexibility, a tolerance value in the range of $\boldsymbol{d}_{ij} \pm 0.5$ Å is set for the upper and lower limits for distance constraints. A maximum tolerance of 10° is set for the backbone dihedral angles.
2. apply geometric constraints to fold the target sequence using the self-correcting distance geometry based program DIAMOD [28, 29].

The input for MPACK is an alignment with the target and a structure file in PDB format for the template. From all generated models, the user can select those with the lowest target function which is a measure of the violation of the geometric constraints during progressive folding of the target sequence. The selected models are then energy minimized to relax steric clashes and optimally place the loop regions and

Table 1. List of Fold-Recognition Servers to Which METASUBMIT Sends the Target Sequence

Server Name	Web site location
BIOSERVER [46]	http://bioserv.cbs.cnrs.fr/HTML_BIO/frame_meta.html
BIOINBGU [47]	http://www.cs.bgu.ac.il/~bioinbgu/form.html
SAM-T99 [48]	http://www.cse.ucsc.edu/research/compbio/HMM-apps/T99-query.html
SUPERFAMILY [17, 49]	http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/hmm.html
FUGUE [50]	http://www-cryst.bioc.cam.ac.uk/~fugue/prfsearch.html
3DPSSM [51, 52]	http://www.sbg.bio.ic.ac.uk/~3dpssm/html/ffrecog_simple.html
GENTHREADER [53]	http://bioinf.cs.ucl.ac.uk/psiform.html

Table 2. The CAFASPSCORER Tool Orders the Alignments from the Various CAFASP Servers According to How Well the Motifs Selected for the Target Match to the Template. The individual motif scores (scaled from 0 to 1) for the best matching templates are given for the 7 PCP-motifs (listed in Table 3) from an alignment of sequences matching T184. Only the highest scoring alignments from each server are shown. 3D-PSSM alignments chosen for the models of the two domains are highlighted.

Server name	PDB code	Template SCOP #	Matching scores for Motifs 1-7 of T184
PDB-Blast	1JFZ_A	a.149.1	0.9462 0.6626 0.6071 0.5125 0.0000 0.0000 0.0000
PDB-Blast	1QU6_A	d.50.1	0.0000 0.0000 0.0000 0.0000 0.2793 0.8173 0.8339
PDB-Blast	1QU6_A	d.50.1	0.0000 0.0000 0.0000 0.0000 0.2866 0.5762 0.7356
PDB-Blast	1DI2_A	d.50.1	0.0000 0.0000 0.0000 0.0000 0.4969 0.8173 0.6916
PDB-Blast	1EKZ_A	d.50.1	0.0000 0.0000 0.0000 0.0000 0.3446 0.9152 0.6311
Sam-T99	1JFZ_A	a.149.1	0.9462 0.6513 0.6071 0.5132 0.3567 0.0000 0.0000
Sam-T99	1DI2_A	d.50.1	0.0000 0.0000 0.0000 0.0000 0.4969 0.8173 0.6680
Sam-T99	1EKZ_A	d.50.1	0.0000 0.0000 0.0000 0.0000 0.3446 0.9152 0.6311
SUPERFAMILY	1JFZ_A	a.149.1	0.9462 0.6513 0.6071 0.5024 0.3283 0.8173 0.8339
FFAS	1DI2_B	d.50.1	0.0000 0.0000 0.0000 0.0000 0.4969 0.8173 0.6916
FFAS	1QU6_A	d.50.1	0.0000 0.0551 0.2607 0.4978 0.2866 0.5186 0.7356
FFAS	1EKZ_A	d.50.1	0.0000 0.0000 0.0000 0.0000 0.3446 0.9152 0.6311
3D-PSSM	**1JFZ_A**	**a.149.1**	**0.9462 0.6513 0.6071 0.5527 0.0000 0.0000 0.0000**
3D-PSSM	**1DI2_A**	**d.50.1**	**0.0000 0.0000 0.0000 0.0000 0.4969 0.8173 0.6916**
3D-PSSM	**1QU6_A**	**d.50.1**	**0.0000 0.0000 0.0000 0.0000 0.3283 0.8173 0.8339**
mGenTHREADER	1I4S_A0	a.149.1	0.9462 0.6626 0.6071 0.5257 0.0000 0.0000 0.0000
mGenTHREADER	1QU6_A1	d.50.1	0.0000 0.0000 0.0000 0.0000 0.3283 0.8173 0.8339
mGenTHREADER	1QU6_A2	d.50.1	0.0000 0.0000 0.0000 0.0000 0.2866 0.5762 0.7356
mGenTHREADER	1DI2_A0	d.50.1	0.0000 0.0000 0.0000 0.0000 0.4969 0.8173 0.6916
FUGUE2	1JFZ_A	a.149.1	0.9462 0.6414 0.6071 0.5369 0.0000 0.0000 0.0000
FUGUE2	1QU6_A	d.50.1	0.0000 0.0000 0.0000 0.0049 0.3283 0.8173 0.8339
FUGUE2	1STU	d.50.1	0.0000 0.0000 0.0000 0.0000 0.3446 0.9152 0.6311
INBGU	1JFZ_A	a.149.1	0.9462 0.6513 0.6071 0.5257 0.0000 0.0000 0.0000
INBGU	1DI2_A	d.50.1	0.0000 0.0000 0.0000 0.0000 0.4969 0.8173 0.6916
INBGU	1STU	d.50.1	0.0000 0.0000 0.0000 0.0000 0.3446 0.9152 0.6311
INBGU	1QU6_A	d.50.1	0.0000 0.0000 0.0000 0.0000 0.3283 0.8173 0.8339
ORNL-PROSPECT	1JFZ_D	a.149.1	0.9462 0.6626 0.6071 0.5257 0.0000 0.0000 0.0000
ORNL-PROSPECT	1JFZ_D	a.149.1	0.9462 0.6626 0.6071 0.0000 0.4969 0.8173 0.1512

side chains using the FANTOM program [30]. Note, however, that due to time constraints we did not include any special method for modeling loop regions, such as those we have used for other examples [4]. The geometry of the model is evaluated with PROCHECK [31]. Additional selection criteria include conformational energy, exposed surface area measured using GETAREA [32], RMSD deviation from the template backbone, and determining the fit of residues in the active site of enzymes or highly conserved regions, using MOLMOL [33].

OVERVIEW OF THE CASP5 MODELS GENERATED USING PCP MOTIFS

Below, we illustrate how PCP-motifs can be used to develop homology models and discriminate functional areas of proteins, using CASP5 targets of potential interest to medical researchers. As with any competition, certain measures are used to determine a ranking of the participants performance. We show here three measures of quality for the models we submitted. The most comprehensive way to

Table 3. ALIGNSCORER Results (Default Conditions) for the Selected Alignments from 3D-PSSM for Target 184 with the Templates 1JFZ (Domain 1) and 1DI2 (Domain 2). PCP-motifs for the template were isolated from an alignment of T184 with other RNase III proteins, using PCPMer.

Motif	TargetMotif	Template 1JFZ match	Score
1	TGINFKNEELLFRALCHSSY	LGYTFKDKSLLEKALTHVSY	0.9462
2	ESNEKLEFLGDAVLELFVCEILYKKYP	HYETLEFLGDALVNFFIVDLLVQYSP	0.6513
3	VGDLARVKSAAAS	EGFLSPLKAYLIS	0.6071
4	LAMVSRKMNLGKFLFLGKGEEKTGGRDRDSILADAF EALLAAIYLDQGYEKIKELF	FNLLAQKLELHKFIRI------KRGKINETI IGDVFEALWAAVYIDSGRDFTRELF	0.5527
5	DYKTALQEIVQ	----------	0.0000
6	KNDG	----	0.0000
7	GKGRTKKEAEKEAARIAYEKL	----------	0.0000

Motif	TargetMotif	Template 1DI2 match	Score
1	TGINFKNEELLFRALCHSSY	----------	0.0000
2	ESNEKLEFLGDAVLELFVCEILYKKYP	----------	0.0000
3	VGDLARVKSAAAS	----------	0.0000
4	LAMVSRKMNLGKFLFLGKGEEKTGGRDRDSILADAFEALLAAIYLDQ GYEKIKELF	---------- ----------	0.0000
5	DYKTALQEIVQ	MPVGSLQELAV	0.4969
6	KNDG	GPPH	0.8173
7	GKGRTKKEAEKEAARIAYEKL	GSGTSKQVAKRVAAEKL LTKF	0.6916

compare models is to use coverage plots, which plot the root mean square deviation (RMSD) between two sequences as a function of the length of the sequence (Fig. (**2**)). A good model should have a large coverage with low RMSD to the experimental structure, i.e. it has a plot close to the X-axis. Our models are similar in quality to the best submitted

Table 4. Models Described in This Review Submitted by Our Group (024) Compared to the Best Model Submitted in CASP5. The PDB code for the template, percentage identity between target and template, FANTOM energy, RMSD and GDT score for our models is given for each target. For comparison, the RMSD, GDT values and the number and name of the group producing the #1 ranked model is given for each target. RMSD and GDT values are taken from the CASP5 web site, http://predictioncenter.llnl.gov/casp5/.

Target	Template	%ID	Energy kcal/mol	RMSD C_α [a]	GDT [b]	Results for the group with highest GDT			
						RMSD C_α [a]	GDT [b]	Group Number	Group Name
T142	1I9Z	27 (73/275)	-862	3.66	65.09	3.46	72.23	035	Lambert-Christophe
T150	1CK2	32 (32/100)	-392	2.15	78.65	1.89	82.56	265	Sasson-Iris
T155	1DHN	33 (40/120)	-1090	1.02	93.38	0.74	98.08	012	ORNL-Prospect
T178	1JCJ	26 (55/211)	-547	6.65	76.03	1.63	84.70	067	Jones
T182	1C24	42 (105/250)	-1790	1.32	91.17	1.28	94.18	329	Dunbrack
T184_1	1JFZ	27 (45/164)	-156	1.88 [c]	67.88	1.90 [c]	74.24	016	Levitt
T184_2	1DI2	34 (24/71)	-733	1.47 [c]	85.07	1.68 [c]	87.85	334	MZ-Brussels
T188	1E01	28 (34/122)	-377	2.29	77.33	2.13	78.50	100	SBI

[a] root-mean-square deviation (RMSD) between the model and experimental structure calculated using a sequence-dependent superposition for C_α

[b] the global distance test (GDT) represents an average of the maximum number of residues that can be superimposed between the experimental structure and the corresponding model under four different distance thresholds (1, 2, 4, and 8 Å) in a sequence-dependent manner: GDT = (GDT_P1+GDT_P2+GDT_P4+GDT_P8)/4, where GDT_Pn is the percent of residues under the distance cutoff of n Å

[c] RMSD between the model and experimental structure calculated on residues that fit under 4.0 Å distance cutoff (GDT_4)

Table 5. PCPMer Identifies Nitrophorin as a Close Relative of the DNase 1 Family. The shared motifs are those within the metal ion binding center and are presumably the site of heme binding. This is the first structure of a DNase 1 superfamily member that binds heme, rather than an isolated metal ion. The sequence of T142 was added to the FASTA format sequence files downloaded from a recent release of the ASTRAL40 database, which contains about 4000 unique sequences representing most of the protein folds in the PDB. This database was then searched with motifs from an alignment of diverse members of the DNase 1 superfamily with the in-house program PCPMer. PCPMer defines motifs in aligned protein sequences as a numerical matrix representing conserved physical chemical properties [9, 25]. The numerical matrices can be used to automatically scan databases for sequence homologs.

Score	PDBfile	protein
1386	1HD7_A	human APE1
1338	1AKO	*E. coli* Xth (bacterial APE1)
1330	2DNJ	Bovine DNase 1
1241	1QGU_A	*Klebsiella pneumoniae* nitrogenase alpha chain
1240	1I9Y_A	inositol 5'-polyphosphate phosphatase
1237	1ACO	bovine aconitase
1224	1HO5	nucleotide 5'-phosphatase
1222	**1NTF**	**T142, *Cimex nitrophorin***
1189	1CJA_A	actin fragment kinase (slime mold)

models. A numerical summary is given in Table 4. For each target, the RMSD and GDT scores for our model is compared to the best model (*i.e.* the model with the highest GDT) using data provided on the CASP5 Web site http://predictioncenter.llnl.gov/casp5/. In all but one case (T178) the RMSD and GDT values are close to those of the best prediction. Even the T178 model is close to the experimental structure (see the coverage plot, Fig. (**2**)), except for the 20 residues in the C-terminal, where we lacked motifs to guide the alignment. Further, no one method yielded the best model for all targets, indicating that it is worthwhile to compare details of the models produced by several groups. The third illustration is perhaps the clearest indication of quality: stereo plots of the models superimposed on the experimental structures (Fig. (**3**)). Below, we highlight some of the details from these models that illustrate the uses of sequence decomposition in analyzing proteins.

T142, Nitrophorin, has PCP-Motifs in Common with the DNase 1 Superfamily

The bedbug, *Cimex lectularius*, secretes nitrophorin (CASP5 target T142), a nitric oxide binding heme protein that is a potent vasodilator, from its salivary glands to facilitate blood sucking. Inhibitors of this enzyme might thus be used to prevent human infestation with these pests. Salivary nitrophorin contains motifs that are common to a superfamily of metal binding proteins, which include DNase 1, DNA repair proteins, apurinic/apyrimidinic endonucleases, and the inositol-5'-polyphosphate phosphatases (IPP) [8, 34]. This is an interesting finding, as the other proteins bind a free metal ion, not a heme group. We were thus interested in seeing how the metal binding site of nitrophorin differed from the non-heme metal binding sites of DNase 1 family members. The template selected by most fold recognition servers was the PDB file 1I9Z, synaptojanin, an IPP that is 25% identical and about 40% similar in the alignment used for modeling. Most of the identical residues were in motifs for the metal centered active site of the DNase 1 family. These include the areas around the conserved aspartate (735 VVWFGDNYRI 745 in 1I9Z, 166 LFWIGDLNVRV 176 in T142), and histidine (837 SDHRPIYATYEA 848; in I9Z; 269 TEHRPVLAKFRV in T142) which are essential for function in the DNase 1 superfamily enzymes. As Fig. (**4**) shows, the structure of these two segments match extremely well in the template, the model, and the determined experimental structure (PDB code 1NTF). Two other motifs known to contribute to the active site of the DNase 1 superfamily also match well in both sequence and structure. Overall, the fact that the two metal sites are so similar suggests that this nitric oxide (NO) transporting protein evolved from enzymes that catalyze phosphorolytic cleavage [8, 9].

As a test to see whether our PCPMer program would recognize T142 as a homolog of the DNase 1 family, we added the T142 sequence to the ~4000 sequences in the ASTRAL40 structural database and searched the database with motifs of the DNase 1 family. The *Cimex* nitrophorin was listed within the first 10 entries, with a PCPMer score that was only slightly below the IPP structure we used as template. Another example of using PCPMer's ability to identify distant homologs of well-established enzyme families in genomic databases is shown below for T188.

T155, DHNA, a Promising Target for Drug Design

Enzymes of pathogenic bacteria are one target for modern drug design. Inhibitors of dihydroneopterin aldolase (DHNA), an enzyme important in the folic acid pathway, are now being considered for treatment of tuberculosis [35]. The DHNA of *Mycobacterium tuberculosis* (Genbank gi3023784), target T155, 133 residues, was 33% identical to the best PDB template, the crystal structure (1.65 Å resolution) of 7,8-dihydroneopterin aldolase from *Staphylococcus aureus*. In this case, the motif selection method indicated the same alignment that was returned by several fold recognition servers, including 3D-PSSM, Fugue, mGenThreader and Inbgu. Our model, which ranked well according to the tabulated data (the experimental structure has not yet been released), had a very low RMSD value to the template (Table 4). This would indicate that DHNA from many pathogenic bacteria should have a similar structure, despite the variation in their sequences, and should be good targets for design of general antibacterial agents.

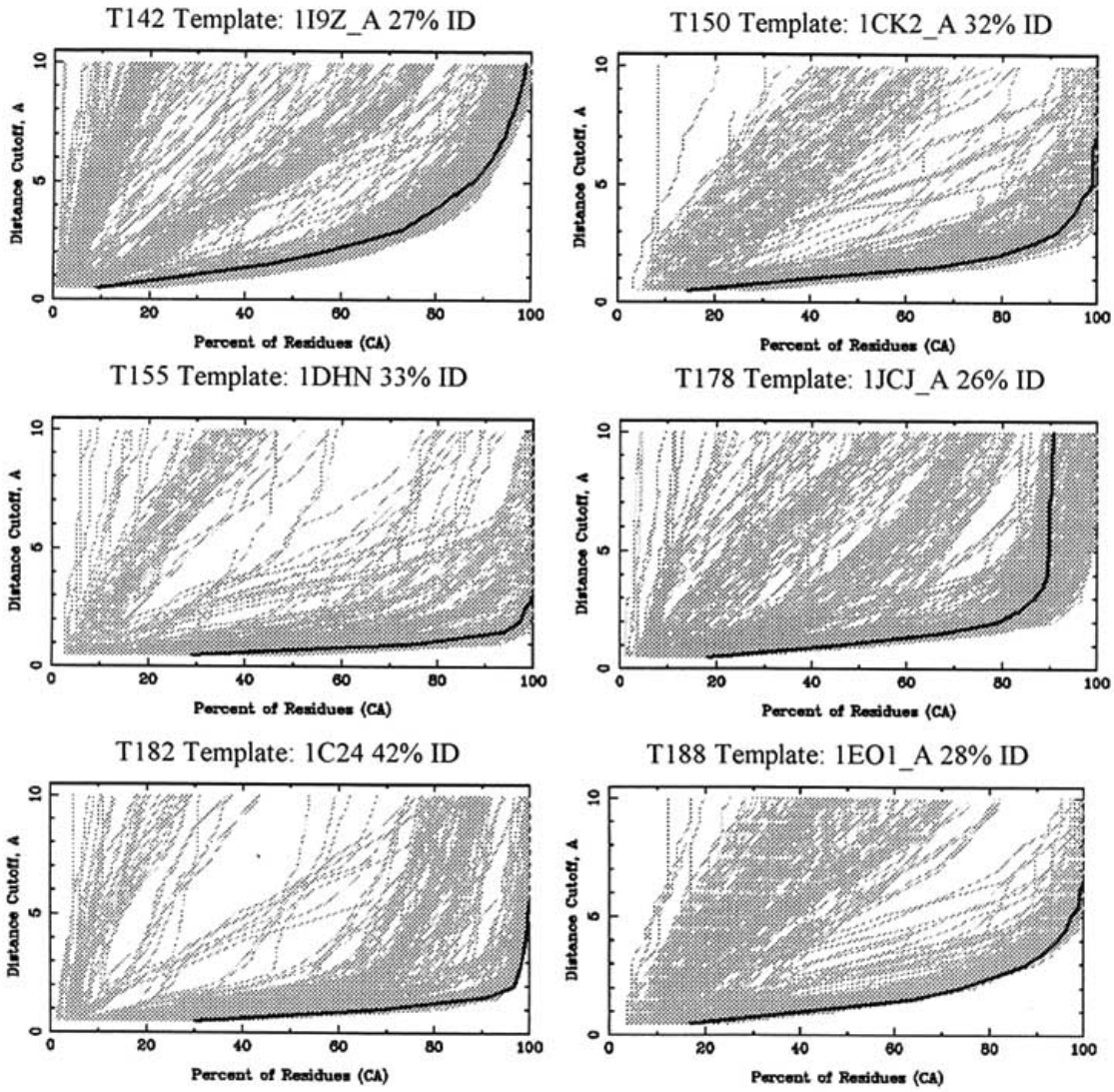


Fig. (2). Coverage or global distance test (GDT) plots for models of CASP5 targets discussed in this review. The X-axis is the percent of residues that falls within the distance cutoff (Y-axis). The best model will be close to the experimental structure over most of the protein sequence. Thus, the nearer the line runs to the X-axis, the better the model. The results for our models are given in bold black lines, the gray lines are the results for models from the other >200 groups that were submitted for the same targets. (The plots were obtained from http://predictioncenter.llnl.gov).

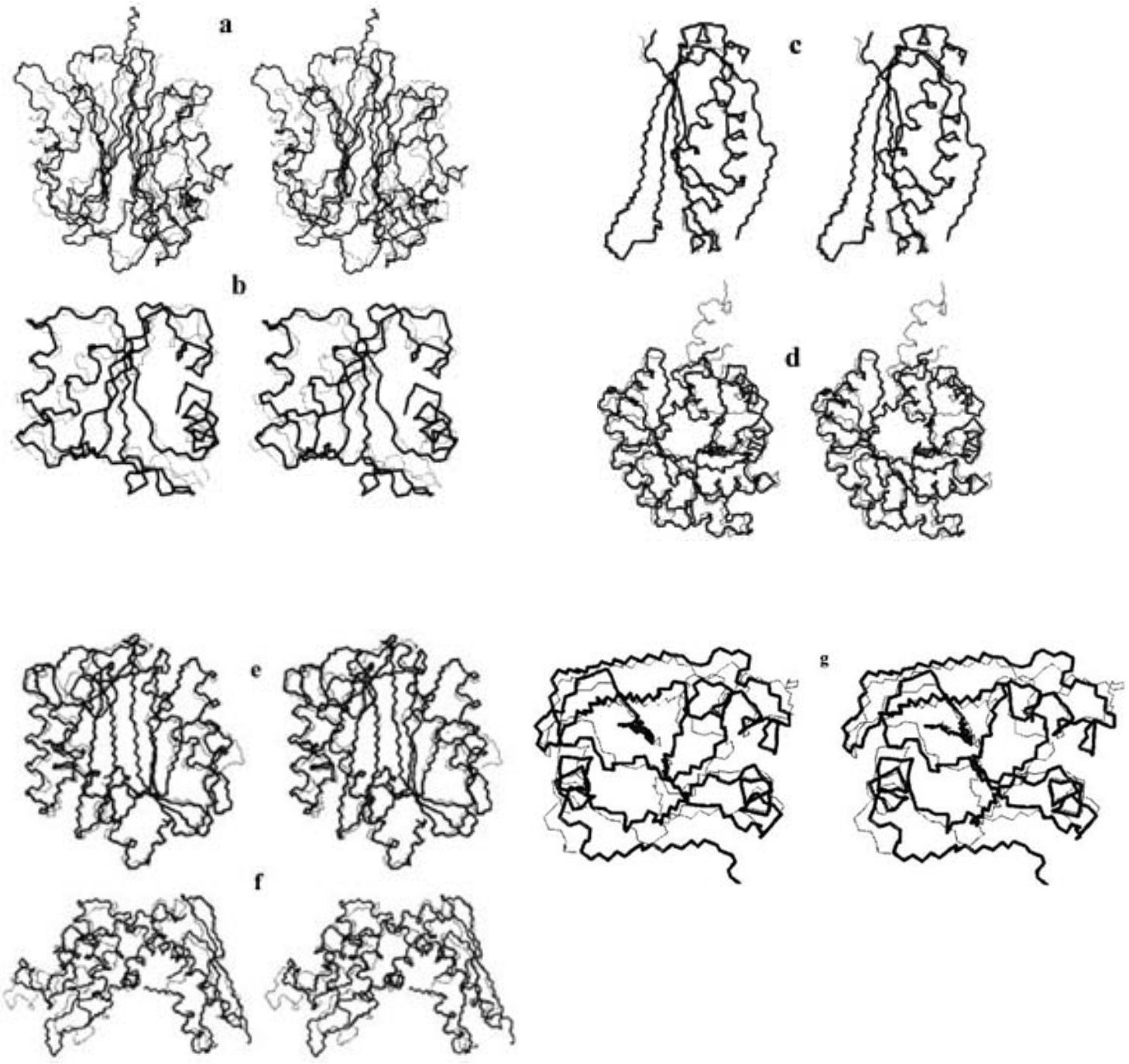


Fig. (3). Stereo plots of the models prepared for the CASP5 targets (heavy black line) discussed in this article superposed on the experimentally determined structures (given as the PDB file name of the released coordinates; light line), which were released after the models were submitted. (a) T142 vs. 1NTF; (b) T150 vs. 1H7M; (c) T155 vs. template 1DHN (structure not yet publicly released); (d) T178 vs. 1MZH; (e) T182 vs. 1O0X; (f) T184 vs. 1O0W (both domains); (g) T188 vs. 1O13. For T184, the two domains of the model were separately matched to the N and C-termini of the crystal structure.

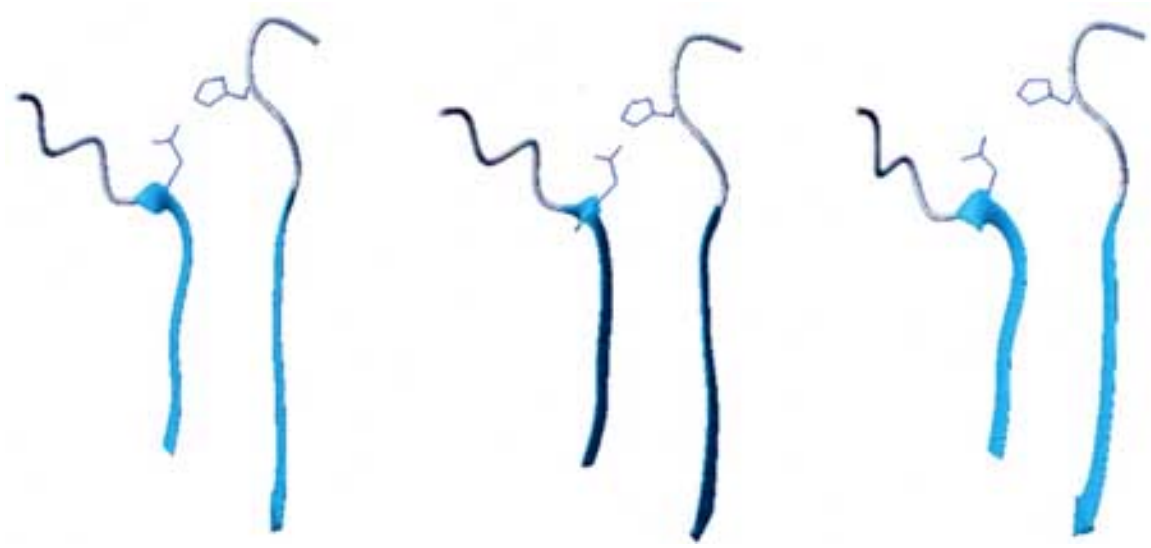


Fig. (4). Molegos, conserved regions of sequence and structure, match well in template, model and experimental structure (left to right) of nitrophorin, T142. This figure shows two isolated molegos, which correspond to APE molegos, 7 and 12 involved in metal ion binding [8]) from the three structures. The two molegos are shown from the template (PDB file 1I9Z, for synaptojanin), our model structure for T142, and the experimental structure of T142 (PDB file 1NTF, which has not yet been released).

Modeling Thermophilic Enzymes, T150, T178 and T182

There is considerable interest in studying the enzymes of thermophiles, as the sequences are often quite different from those of bacteria that grow at lower temperatures [36]. These targets illustrate that thermophilic enzymes can be accurately modeled on homologous enzyme targets from mesophiles. Target T150 was the 102 residue ribosomal protein L30e from *Thermococcus celer* (SwissProt P29160). Our MPACK model used the yeast *Saccharomyces cerevisiae* ribosomal protein L30 (PDB code 1CK9 for 20 NMR structures and 1CK2 for the minimized average structure) and the consensus alignment from Fugue, mGenThreader and Inbgu. Target T182, the 250 residue protein methionine aminopeptidase (E.C. 3.4.11.18) from *Thermotoga maritima* was modeled on the methionine aminopeptidase from *Escherichia coli* (PDB code 1C24_A) using an alignment with 42% sequence identity between the target and the template. The low RMSD between the models and the experimental structures indicate thermophilic and mesophilic proteins have similar overall structures.

We used a motif-based approach to model target T178, the 219 residue deoxyribose-phosphate aldolase (DERA) of the hyperthermophilic bacterium *Aquifex aeolicus*, as there were many possible alignments with the selected template. DERA catalyzes a reversible aldol reaction between acetaldehyde, and the acceptor substrate, D-glyceraldehyde-3-phosphate, to generate D-2-deoxyribose-5-phosphate [37]. The reverse reaction of this enzyme allows microorganisms to use nucleotides and even DNA in their growth medium, by bringing pentose-5-phosphate into the glycolytic cycle [38]. To extract PCP motifs, we used an alignment of 26 sequences for aldolase family members, primarily from extremophilic and pathogenic bacteria. While the overall identity to the template suggested by most fold recognition servers, mesophilic *E. coli* DERA, (PDB file 1JCJ_A), was only 26%, the 8 motifs could be matched well in two alignments from different servers. Our model for the first 198 residues of T178, based on the mGenthreader alignment, was very good, with a bb-RMSD of 1.58Å to the experimental structure. However, the model deviated from the experimental structure in the 21 residues at the C-terminus where there are no motifs (Fig. (**2**)). Our model is especially precise in the active center, and the lysine (K150), which forms a Schiff base with the donor acetaldehyde, is in the same orientation as in the crystal structure.

T184: a 2-Domain Target, RNaseIII

Another thermophilic protein, T184, was the probable RNase III of *Thermobacteria maritime*, which is a two domain protein [39]. The second column of Table 3 lists the sequences of the 7 PCP motifs generated for T184 and related RNAse III protein sequences. The matrix of these motifs was used to rank the target-template alignments provided by the CAFASP fold recognition servers (Table 2). The selected template for the larger nuclease domain and metal binding site was the RNase III of the ultrathermophile

Aquifex aeolicus [40]. For the C-terminus, the template was the ds-RNA binding domain from a mesophilic protein, PDB file 1DI2, from *Xenopus laevis.*

The metal ion binding center of the model matches the final crystal structure of the target extremely well. In particular, the motifs, EKLEFLGDAV, EVGDLA, DAFEAL (underlined indicate absolutely conserved residues) that are common to RNase III enzymes from humans to *Drosophila* to bacteria, match especially well in the model to the determined structure.

The crystal structure of T184 emphasized the importance of including both domains in a single crystal structure, as the ds-RNA binding domain is indeed close to the previously determined metal center of the nuclease domain. While we could have attempted to superimpose the ds-RNA binding domain on the active site region, this would have required additional biological data and considerably more time to make the model than is possible given the time constraints of CASP5. The presence of the C-terminal domain does not greatly affect the active site described for the N-terminal half of the *A. aeolicus* protein [40]. This means that our decision to submit the model as two independent domains, and leave their relative orientation open, was justified.

T188: Using PCPMer to Identify Possible Functions

This target presented a problem that is indicative of the direction of structural biology and genomics. That is, we have a sequence, similar to ones found in many organisms, and we have a structure, but we still do not know the function of the protein. Target T188 was a 124 residue hypothetical protein from *Thermotoga maritima*. All the fold-recognition servers scored another hypothetical protein, MTH1175 from methanobacteria (PDB code 1E01) as the best template, meaning that we were unable to suggest a function for this protein based on the templates returned from the fold recognition servers. Our model ranked fourth among the models submitted in CASP5 for this target. Nearly 80% of the residues were in structurally equivalent positions in the model and the experimental structure.

PCPMer was then used to suggest possible functions for this protein, by determining which other proteins contained PCP-motifs that were common to T188 and its bacterial relatives. A PSIBLAST search with E-value cutoff of 0.005, which converged in 3 iterations, yielded 15 sequences similar to T188, all of which lacked a functional annotation. These sequences were aligned with CLUSTALW and seven PCP-motifs were generated, using MASIA (G-cutoff 2, L-cutoff 4, R-cutoff 1.25). The PCP-motifs were used to screen the ASTRAL40 database of representative known structures to detect proteins with similar motifs (Table 6). PCPMer ranked the template we had chosen, 1EO1, highest of the PDB files in the sequence list. The next three proteins, with similar PCP-scores, are from different SCOP families and have different activities. However, they all contain a Fe-S center, suggesting that T188 must contain motifs that correlate with the metal binding of these proteins. PCPMer results suggest several experimentally testable leads to determine the function of the hypothetical protein encoded by T188.

CONCLUSIONS

The CASP5 Results Validate our Models for Use in Allergy and Vaccine Studies

These results from our CASP5 participation indicate that our modeling approach generates useful and high quality models. The models are especially accurate in regions where we could detect conserved motifs and active sites. This validates the modeling methods used in our other collaborative projects of proteins of high medical interest [1-4, 41-43]. We are now preparing models of all known allergenic proteins that will be made available via our

Table 6. PCPMer Results for Scanning the ASTRAL40 Database with PCP-Motifs Derived from Aligned Sequences Similar to T188, a Hypothetical Protein. See the heading of Table 5 for more details about the methodology. The highest scoring protein, 1EO1, is the modeling template. The next three finds all contain a FeS cluster in their active center. The fourth find, carboxylesterase, is an α/β hydrolase with broad substrate specificity. A structural filter will be applied to these finds to determine which could be functional homologs of T188.

Parameters used: R 1.25, G cutoff 2, L cutoff 4.

List of Motifs:

```
#MOTIF : 1:   1 MIIAIPVSENRGKDSPI 17
#MOTIF : 2:  21 FGRAPYFAFVK 31
#MOTIF : 3:  64 GAELVI 69
#MOTIF : 4:  73 IGRRA 77
#MOTIF : 5:  81 FEAMGVKVIKGASGTVEEVVN 101
#MOTIF : 6: 114 EVHDHHHHEH 123
```

Top hits from PCPMer

Name of the protein	PDB code	SCOP ID	Scores in bits
Hypothetical protein *MTH1175 (Archaea, *Methanobacterium thermoautotrophicum*)	1E01_A	c.55.5.1	417.42
Formate dehydrogenase H (*Escherichia coli*)	1AA6_2	c.81.1.1	397.57
Ferric enterobactin receptor FepA (*Escherichia coli*)	1FEP_A	f.4.3.3	397.52
Xanthine oxidase C-terminal domain (*Bos Taurus*)	1FO4_A	d.133.1.1	394.72
Carboxylesterase (*Pseudomonas fluorescens*)	1AUO_A	c.69.1.14	388.56

structural database of allergenic proteins (SDAP) webserver [44, 45]. The results detailed here for a wide selection of different proteins, and varying degrees of modeling difficulty, show how far we have come in being able to predict the structure of a protein from its sequence and the structure of a homolog.

PCPMer, a New Method for Protein Analysis

Our results also illustrate the use of physical-chemical property motifs to select templates for modeling and identify functionally important areas in novel proteins. Data from sequence decomposition, as was done for example for T188, can suggest function for unknown genomic sequences and in identifying distant relatives of known proteins [8, 9]. We have begun using this method to compare virus sequences, to reveal areas for drug and vaccine design. The PCPMer method and associated tools provide an alternative way to analyze protein sequences that will have wide uses in our future work.

ACKNOWLEDGEMENTS

This work was supported by grants from the U.S. Department of Energy (DE-FG-00ER63041), the John Sealy Memorial Endowment Fund (#2535-01), the U.S. Food and Drug Administration (FDA-U-002249-1).

REFERENCES

[1] Hasan, R. J.; Pawelczyk, E.; Urvil, P. T.; Venkatarajan, M. S.; Goluszko, P.; Kur, J.; Selvarangan, R.; Nowicki, S.; Braun, W. A.; Nowicki, B. J. *Infect. Immun.*, **2002**, *70*, 4485-4493.

[2] Mumenthaler, C.; Schneider, U.; Buchholz, C. J.; Koller, D.; Braun, W.; Cattaneo, R. *Protein Sci.*, **1997**, *6*, 588-597.

[3] Soman, K.; Midoro-Horiuti, T.; Ferreon, J.; Goldblum, R.; Brooks, E.; Kurosky, A.; Braun, W.; Schein, C. H. *Biophys. J.*, **2000**, *79*, 1601-1609.

[4] Murtazina, D.; Puchkaev, A. V.; Schein, C. H.; Oezguen, N.; Braun, W.; Nanavati, A.; Pikuleva, I. A. *J. Biol. Chem.*, **2002**, *277*, 37582-37589.

[5] Casasnovas, J.; Larvie, M.; Stehle, T. *EMBO Journal*, **1999**, *18*, 2911-2922.

[6] Moult, J.; Fidelis, K.; Zemla, A.; Hubbard, T. CASP5. Fifth Meeting of the Critical Assessment of Techniques for Protein Structure Prediction **2002**.

[7] Tramontano, A.; Leplae, R.; Morea, V. *Proteins*, **2001**, 22-38.

[8] Schein, C. H.; Ozgun, N.; Izumi, T.; Braun, W. *B M C Bioinformatics*, **2002**, *3*, art. no.-37.

[9] Mathura, V. S.; Schein, C. H.; Braun, W. *Bioinformatics*, **2003**, *19*, 1381-1390.

[10] Bowie, J. U.; Luthy, R.; Eisenberg, D. *Science*, **1991**, *253*, 164-170.

[11] Gribskov, M.; Veretnik, S. In *Methods in Enzymology*; Academic Press: San Diego, **1996**; Vol. *266*, pp. 198-212.

[12] Mehta, P.; Argos, P.; Barbour, A.; Christen, P. *Proteins*, **1999**, *35*, 387-400.

[13] Schaffer, A. A.; Wolf, Y. I.; Ponting, C. P.; Koonin, E. V.; Aravind, L.; Altschul, S. F. *Bioinformatics*, **1999**, *15*, 1000-1011.

[14] Yona, G.; Levitt, M. *J. Mol. Biol.*, **2002**, *315*, 1257-1275.

[15] Rychlewski, L.; Jaroszewski, L.; Li, W.; Godzik, A. *Protein Sci*, **2000**, *9*, 232-241.

[16] Eddy, S. R. *Bioinformatics*, **1998**, *14*, 755-63.

[17] Gough, J.; Chothia, C. *Nucleic Acids Res.*, **2002**, *30*, 268-272.

[18] Martelli, P. L.; Fariselli, P.; Krogh, A.; Casadio, R. *Bioinformatics*, **2002**, *18 Suppl 1*, S46-53.

[19] Falquet, L.; Pagni, M.; Bucher, P.; Hulo, N.; Sigrist, C. J. A.; Hofmann, K.; Bairoch, A. *Nucleic Acids Res.*, **2002**, *30*, 235-238.

[20] Henikoff, J. G.; Greene, E. A.; Pietrokovski, S.; Henikoff, S. *Nucleic Acids Res.*, **2000**, *28*, 228-230.

[21] Zhu, H.; Schein, C. H.; Braun, W. *Bioinformatics*, **2000**, *16*, 950-951.

[22] Schaffer, A. A.; Aravind, L.; Madden, T. L.; Shavirin, S.; Spouge, J. L.; Wolf, Y. I.; Koonin, E. V.; Altschul, S. F. *Nucleic Acids Res.*, **2001**, *29*, 2994-3005.

[23] Altschul, S.; Madden, T.; Schaffer, A.; Zhang, J. H.; Zhang, Z.; Miller, W.; Lipman, D. *Faseb J.*, **1998**, *12*, 102.

[24] Thompson, J.; Higgins, D.; Gibson, T. *Nucleic Acids Res.*, **1994**, *22*, 4673-4680.

[25] Venkatarajan, M. S.; Braun, W. *J. Mol. Model.*, **2001**, *7*, 445-453.

[26] Lo Conte, L.; Brenner, S. E.; Hubbard, T. J. P.; Chothia, C.; Murzin, A. G. *Nucleic Acids Res.*, **2002**, *30*, 264-267.

[27] Soman, K. V.; Schein, C. H.; Zhu, H.; Braun, W. In *Nuclease Methods and Protocols, Methods in Molecular Biology*, C. H. Schein, ed.; Humana Press: Totowa, **2001**; Vol. *160*, pp. 263-286.

[28] Sanner, M.; Widmer, A.; Senn, H.; Braun, W. *J. Comput.-Aided Mol. Des.*, **1989**, *3*, 195-210.

[29] Mumenthaler, C.; Braun, W. *J. Mol. Biol.*, **1995**, *254*, 465-480.

[30] Schaumann, T.; Braun, W.; Wuthrich, K. *Biopolymers*, **1990**, *29*, 679-694.

[31] Laskowski, R.; Rullmann, J.; MacArthur, M.; Kaptein, R. *J. Biomol. NMR*, **1996**, *8*, 477-486.

[32] Fraczkiewicz, R.; Braun, W. *J. Comput. Chem.*, **1998**, *19*, 319-333.

[33] Koradi, R.; Billeter, M.; Wuthrich, K. *J. Mol. Graph.*, **1996**, *14*, 51-55.

[34] Valenzuela, J. G.; Ribeiro, J. M. C. *J. Exp. Biol.*, **1998**, *201*, 2659-2664.

[35] Deng, H.; Callender, R.; Dale, G. E. *J. Biol. Chem.*, **2000**, *275*, 30139-30143.

[36] Galperin, M. Y.; Aravind, L.; Koonin, E. V. *FEMS Microbiol. Lett.*, **2000**, *183*, 259-264.

[37] Heine, A.; DeSantis, G.; Luz, J. G.; Mitchell, M.; Wong, C. H.; Wilson, I. A. *Science*, **2001**, *294*, 369-374.

[38] Sgarrella, F.; Poddie, F. P. A.; Meloni, M. A.; Sciola, L.; Pippia, P.; Tozzi, M. G. *Comp. Biochem. Physiol. B-Biochem. Mol. Biol.*, **1997**, *117*, 253-257.

[39] Ramos, A.; Pastore, A. *Methods Mol. Biol.*, **2001**, *160*, 237-248.

[40] Blaszczyk, J.; Tropea, J. E.; Bubunenko, M.; Routzahn, K. M.; Waugh, D. S.; Court, D. L.; Ji, X. H. *Structure*, **2001**, *9*, 1225-1236.

[41] Schein, C. H.; Nagle, G. T.; Page, J. S.; Sweedler, J. V.; Xu, Y.; Painter, S. D.; Braun, W. *Biophys. J.*, **2001**, *81*, 463-472.

[42] Mathura, V. S.; Soman, K. V.; Varma, T. K.; Braun, W. *J. Mol. Model.,* **2003**, *9*, 298-303.

[43] Ivanciuc, O.; Mathura, V.; Midoro-Horiuti, T.; Braun, W.; Goldblum, R.; Schein, C. H. *J. Agric. Food Chem.*, **2003**, *51*, 4830-4837.

[44] Ivanciuc, O.; Schein, C. H.; Braun, W. *Bioinformatics*, **2002**, *18*, 1358-1364.

[45] Ivanciuc, O.; Schein, C. H.; Braun, W. *Nucleic Acids Res.*, **2003**, *31*, 359-362.

[46] Douguet, D.; Labesse, G. *Bioinformatics*, **2001**, *17*, 752-753.

[47] Fischer, D. *Pac. Symp. Biocomputing*, **2000**, 119-130.

[48] Karplus, K.; Sjolander, K.; Barrett, C.; Hughey, R. *Bioinformatics*, **1998**, *14*, 846-856.

[49] Gough, J.; Karplus, K.; Hughey, R.; Chothia, C. *J. Mol. Biol.*, **2001**, *313*, 903-919.

[50] Shi, J.; Blundell, T. L.; Mizuguchi, K. *J. Mol. Biol.*, **2001**, *310*, 243-257.

[51] Kelley, L. A.; MacCallum, R. M.; Sternberg, M. J. *J. Mol. Biol.*, **2000**, *299*, 499-520.

[52] Bates, P. A.; Kelley, L. A.; MacCallum, R. M.; Sternberg, M. J. *Proteins*, **2001**, *Suppl 5*, 39-46.

[53] McGuffin, L. J.; Bryson, K.; Jones, D. T. *Bioinformatics*, **2000**, *16*, 404-405.